Docket No.: M4065.0381/P381

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICATION FOR U.S. LETTERS PATENT

Title:

TECHNIQUE TO MITIGATE SHORT CHANNEL EFFECTS WITH VERTICAL GATE TRANSISTOR WITH DIFFERENT GATE MATERIALS

Inventors:

Leonard Forbes

Luan C. Tran

Kie Y. Ahn

Dickstein Shapiro Morin & Oshinsky LLP
2101 L Street NW
Washington, DC 20037-1526
(202) 785-9700

TECHNIQUE TO MITIGATE SHORT CHANNEL EFFECTS WITH VERTICAL GATE TRANSISTOR WITH DIFFERENT GATE MATERIALS

FIELD OF THE INVENTION

This invention relates to the field of semiconductor transistors that are scaled down to sub-micron sizes.

BACKGROUND OF THE INVENTION

There is ever-present pressure in the semiconductor industry to develop smaller and more highly integrated devices. As the industry standard approaches smaller and smaller scaled devices, problems with further advancement are presented and it becomes more difficult to produce sub-micron devices that can perform as desired.

As MOSFET are scaled to deep sub-micron dimensions it becomes increasingly difficult to maintain an acceptable aspect ratio, as shown in Fig. 1a. Fig. 1a shows a representational illustration of a MOSFET having a polysilicon gate 3 over a substrate 5, with the two being separated by a gate oxide 7. Source and drain regions 9 of the substrate are on either side of the gate structure, forming a transistor. The aspect ratio equation represents the spatial relationship between the elementary parts of a MOSFET device, specifically between the distance between the source/drain areas defining the effective gate length (L), the width of the depletion region (W_d), the depths of the source/drain areas (x_j), and the gate oxide thickness (t_{ox}). Detrimental short-channel effects occur when the gate length (L) is reduced by the same order as the width of the depletion region (W_d). In current trends, not only are the gate oxide thicknesses scaled to under 5 nm (50 Å) dimensions as the channel lengths are shortened to sub-micron sizes, but also, the

depletion widths (synonymous with W_d) and source/drain junction depths (x_j) must be scaled to smaller dimensions as well. The depletion region width (or space charge) (W_d) are made smaller by increasing the substrate or channel dopings. However, it is extremely difficult to scale junction depths to under 100 nm dimensions because these are doped by ion implantation and thermally activated.

Related to aspect ratio are short channel effects, which are highly dependent on the channel length. For shorter channel devices (channel lengths below 2 μm) a series of effects arise that result in deviations from the predictable performance of larger scaled devices. Short channel effects impact threshold voltage, subthreshold currents, and I-V behavior beyond threshold. Techniques have been developed for avoiding short channel effects in MOSFETs, such as the "straddle gate" transistor shown in Fig. 1b. Such a structure utilizes thinner gate oxides 11 under the gate sidewall spacers 21 to allow the regions to turn-on easier and at lower voltages. A thicker gate oxide 15 is provided beneath the gate 17. These thinner regions produce a "virtual" source/drain junction 19 with minimal junction depth. The problem with such structures is that gate oxides are already approaching theoretical minimal values, therefore, regions of even thinner gate oxides pose reliability risks. It would be beneficial to devise a semiconductor having an acceptable aspect ratio, where the channel length is large enough when the device is "off" to avoid short channel effects and undesired shorting of the device, and where the device channel is short enough when the device is "on" to allow for the fastest operation possible.

SUMMARY OF THE INVENTION

This invention relates to a process of forming a transistor having three adjacent gate electrodes and the resulting transistor. In forming such a transistor it is possible to mitigate short channel effects as MOSFET structures are scaled down to sub-micron sizes.

This transistor fabrication process can utilize different materials for the gate electrodes so that the workfunctions of the three gate electrodes can be tailored to be different. The three gate electrodes can be connected by a single conducting line and all three are positioned over a single channel and operate as a single gate having a pair of outer gate regions and an inner gate region. This allows for use with higher source and drain voltages. These devices provide for higher performance, using a standard or scaled down transistor surface area, than can be achieved with conventional transistor structure. They have smaller effective channel lengths when "on," and consequently, faster speeds are achievable. The devices have longer channel lengths when "off," thereby mitigating short channel effects.

In an alternative arrangement the two side gate electrodes can be independently biased to a fixed voltage to turn on portions of the channel regions over source/drain extensions and the inner gate can subsequently turn on a portion of the channel region between the source/drain regions.

These and other features and advantages of the invention will be more clearly understood from the following detailed description of the invention which is provided in connection with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1a is an illustration of a typical semiconductor transistor and the relationship of its dimensions to the aspect ratio equation;

Fig. 1b is an illustration of a "straddle gate" transistor;

Fig. 2 is an illustration of a semiconductor device transistor in accordance with the invention;

Fig. 2a is an illustration of an alternative embodiment of the semiconductor device of Fig. 2.

Fig. 3a and 3b illustrate the principles of workfunction as it relates to the different areas of a semiconductor device having low and high V_t in accordance with the invention;

Figs. 4-9 show a cross section of a semiconductor device during successive steps of processing in accordance with the invention;

Fig. 10a and 10b illustrate the principles of workfunction as it relates to the different materials of the semiconductor device in accordance with the invention; and

Fig. 11 is an illustration of a processor system utilizing a semiconductor device in accordance with the invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

In the following detailed description, reference is made to various specific embodiments of the invention. These embodiments are described with sufficient detail to enable those skilled in the art to practice the invention, and it is to be understood that other embodiments may be employed, and that structural and electrical changes may be made without departing from the spirit or scope of the present invention.

In the following discussion the terms "wafer" and "substrate" are used interchangeably and are to be understood to refer to any type of semiconductor substrate, including silicon, silicon-on-insulator (SOI), and silicon-on-sapphire (SOS) technology, and other semiconductor structures. Furthermore, references to a "wafer" or "substrate" in the following description, do not exclude previous processing steps utilized to form regions or junctions in or on the base semiconductor structure or foundation.

No particular order is required for the method steps described below, with the exception of those logically requiring the results of prior steps. Accordingly, while many of

the steps discussed below are discussed as being performed in an exemplary order, this order may be altered.

This invention relates to a process of forming a transistor with three adjacent gate electrodes and the resulting transistor. The transistor mitigates short channel effects as MOSFET structures are scaled down to sub-micron sizes and increases the performance of these devices. The transistor can be fabricated with different gate materials so that the workfunctions of the three gate electrodes can be tailored, thereby improving device behavior when "on" and "off." The three gate electrodes can be connected by a single conducing line and all three gate electrodes are provided over a single channel and operate as a single gate having a pair of outer gate regions and an inner gate region. Alternatively, the two side gate electrodes can be independently biased to a fixed voltage to turn on portions of the channel adjacent the source/drain regions, creating extensions, and the inner gate electrode can be turned on for the remainder of the channel by an independent voltage driving source. In such a configuration, the series resistance of the source/drain extensions can be adjusted for by controlling the inner or outer gate voltages.

Referring now to the drawings, where like elements are designated by like reference numerals, a transistor structure formed in accordance with the invention is shown in Fig. 2. For exemplary purposes, the transistor is shown as part of a DRAM memory cell having a bit line 62 and bit line plug 60, and a capacitor plug 64 and a capacitor 66, shown in dashed lines. However, the transistor is not limited to such a use and may be used for other memories (e.g., SRAM, Flash, etc.), general logic, processor, and ASIC applications . The transistor includes a gate dielectric 12 and 13 over a substrate 10, and three thin vertical gate structures 40, 42a and 42b. The outer two gate structures 42a and 42b are of one conductive type material, preferably N+ type polysilicon; and the center gate structure 40 is a different conductive type material, preferably P+ type polysilicon. All three gate

electrodes can be doped by implant using As or P ions for the N+ conductivity type polysilicon and BF_2 or B for the P+ type polysilicon. If the gate materials are deposited at different steps, then in-situ doped polysilicon can be used, using PH_3 for N+, or Diborane for P+ polysilicon. Alternatively, the gate electrode arrangement may be just the opposite where the channel is to be a p-channel and the device is fabricated over an N-well. In such a scenario, gate electrode 40 is N+ polysilicon and the outer gate electrodes 42a and 42b are P+ polysilicon. Additionally, instead of using only doped polysilicon for the gate electrodes 40, 42a and 42b, it is possible to use a metal gate electrode (e.g., W, TiN, TaN, or Mo) where the middle gate 40 will have a higher workfunction than the outer gate electrodes 42a and 42b.

In the first arrangement, the N+ gate electrodes 42a and 42b are separated by a thin dielectric layer 22 from the P+ gate electrode 40, and the tops of all three gate electrodes 40, 42a, and 42b are connected by a single conductive cap 26, which is preferably doped polysilicon, but can alternatively be self-aligned silicide, $TiSi_2$, or $CoSi_2$. This device is effectively the same size (or smaller) and overall shape as a standard DRAM type transistor gate and may be utilized in virtually any semiconductor transistor device.

The inner 40 and outer 42a, 42b gate electrodes can be formed to have different workfunctions (by choice of different conductivity types and/or material types) so that upon turning on, source/drain 32 extensions 46, which consist of virtual source/drain junctions with minimal junction depth, are created by inversion of the transistor channel region below the outer gate structures 42a and 42b, thereby shortening the effective channel length of the device and allowing for faster operation. These virtual source/drain junctions 46 are not present when the device is not operating, so the actual channel length is long enough to avoid undesirable short channel effects.

There can be over a one-volt difference in the workfunction between the gate electrodes of different types. The threshold voltage (V_t) equation has four terms, the Fermi potential ($2\varphi_f$), the bulk charge (Q_B), the oxide charge (Q_{ox}), and the workfunction difference (φ_{ms}). The equation for V_t can be written as follows:

$$V_t = +|2\varphi_f| + |Q_B/C_{ox}| - |Q_{ox}/C_{ox}| + \varphi_{ms}$$

The Fermi potential is dependent on channel doping and increases with increased doping. The bulk charge behaves the same way, but in a square root relationship. C_{ox} is the normalized gate dielectric capacitance and increases as the gate dielectric thickness is reduced. The oxide charge is a function of gate dielectric processing and includes a fixed and interface charge. The workfunction difference is dependent on the gate material and is weakly dependent on the Fermi level of the substrate.

The potential (workfunction) in the gate material is a characteristic property of the material itself. In reference to Figs. 3a and 3b, the use of different materials for the three transistor gate electrodes 40, 42a, and 42b such as those materials described herein, can tailor the transistor's V_t under the different vertical gate electrodes 40, 42a, and 42b by utilizing the inherent workfunctions of those differing materials. This allows the short channel effects to be mitigated by enabling the virtual source/drain 32 extensions 46 (junctions) to be created only when transistor is "on," thereby shortening the transistor's effective channel lengths resulting in faster device performance characteristics. However, as stated, when "off," the device channel length can be large enough to avoid short channel effects. Fig. 3b relates to the channel 48 region under the central gate 40 of the Fig 2 transistor. It illustrates the workfunction difference of the P+ poly gate 40 compared to a P-type substrate 10, where E_f is the Fermi level energy (inside band gap), E_v is the valence band energy, and E_c is the conduction energy. The P+ poly gate 40 results in a more

positive workfunction relative to the substrate and thus a higher relative V_t. As illustrated by Fig. 3a, which relates to the source/drain 32 extensions 46 under the outer gate electrodes 42a and 42b of the Fig. 2 transistor, the N+ poly results in a more negative workfunction relative to the substrate, and thus, a lower relative V_t.

In accordance with the invention, changing the gate materials of the various gate electrodes 40, 42a, and 42b will change the band gap energy. This results in differences in the workfunctions between the outer 42a, 42b and center 40 gate electrodes, and as a consequence different threshold voltages. It is the tailoring of the three gate electrodes' 40, 42a, and 42b threshold voltages that allows for the forming of the virtual source/drain 32 extensions 46. The N+ gate electrodes 42a and 42b have a more negative workfunction and have low V_t, and therefore, tend to be inverted or conduct at near zero gate bias and need no V_t adjustments by ion implantation; they can turn on as soon as appropriate gate potential is applied. The P+ gate 40 has a more positive workfunction, resulting in a V_t that can be one volt or more positive and require more voltage to turn on than the N+ gate electrodes 42a and 42b, thus they will turn on after the N+ outer gate electrodes 42a and 42b. This results in the ability to fabricate faster, scaled down devices because of such devices' ability to avoid the short channel effects that would normally occur due to the reduced channel 48 length while still having a shortened effective channel length. The outer N+ gate electrodes 42a and 42b should each occupy no more than about 10 to 33% of the total channel 48 length, preferably no more than 25% each. As illustrated in Fig. 3b, almost any workfunction difference may be obtained by using a wide band gap gate material, with low electron affinity and doping the gate material to be P type or N type.

The Fig. 2 transistor 44 works as follows. As gate voltage is applied to the conductive cap 26 and thus to the three vertical gate electrodes 40, 42a and 42b, the channel 48 regions under the N+ polysilicon gate electrodes 42a and 42b will become

inverted and these outer gate electrodes will turn on first. By the time a sufficient threshold voltage for the N+ gate electrodes 42a and 42b (effectively any applied voltage) is applied, there will be conductive regions under the N+ gate electrodes, which act as "virtual" S/D (source/drain) extensions 46 (of the actually formed source and drain to which they are adjacent) with minimal junction depth. When enough voltage is applied the P+ gate 40 will start to turn on and normal transistor action will follow. The conductivity under the transistor 44 (channel 48 region) can be the same throughout. The V_t of the P+ gate 40 can be around 0.3 volts, which is appropriate for deep sub-micron dimensioned devices.

As illustrated in Fig. 2a, as an alternative embodiment, the central gate electrode 40 and the outer two gate electrodes 42a and 42b can be independently biased. The overlying polysilicon cap 26 from Fig. 2, can alternatively be formed as one conductive cap 26a in electrical contact with the central gate electrode 40 and a conductive ring cap 26b in contact with the outer two gate electrodes 42b. In this way, separate and tailored voltages can be applied independently to the three adjacent gate electrodes 40, 42a, and 42b, further tailoring the device. These separate conductive caps 26a and 26b are electrically insulated from one another. Alternatively, the outer gate electrodes 42a and 42b can be connected to a sidewall electrical contact (not shown) instead of the conductive ring cap 26b.

As illustrated in Fig. 4 to Fig. 9, the transistor 44 can be formed as follows. Referring to Fig. 4, a semiconductor substrate 10 is provided. A sheet ion implantation and V_t adjustment is performed on substrate 10. LOCOS or STI (Shallow Trench Isolation) can be performed to form FOX (Field Oxide) regions 14 to isolate the devices and a sacrificial oxide can be grown over the substrate to correct defects caused by the STI. If a sacrificial oxide is grown, it is removed prior to further processing.

Next, referring to Fig. 4, after a wafer surface cleaning by a standard RCA clean, a thin gate dielectric 12 is grown over the substrate 10, by, for example, thermal oxidation. Thermal nitridation, which produces a self limited silicon nitride barrier of up to about 20Å or 2.0 nm, can be used to harden the gate dielectric 12 when a P+ center gate electrode 40 is utilized. RTP (Rapid Thermal Nitridation) in NH_3 or Remote Plasma Nitridation (RPN) is sufficient for this purpose. The nitridation will produce a good diffusion barrier for the gate dielectric 12, which can prevent Boron penetration from P+ doped polysilicon (or amorphous silicon). The gate dielectric 12 should be as thin as possible (e.g., 1.0-3.0 nm) to still maintain standard device functioning, as is known in the art.

Next, referring to Fig. 5, a P+ doped polysilicon layer is formed over the wafer and the gate dielectric 12. This layer can alternatively be amorphous silicon. Over the P+ polysilicon layer is deposited a layer of conductive material, such as Tungsten (W), and a protective cap. These layers (polysilicon, conductive material and protective cap) are patterned and etched using the gate dielectric as a stop to form gate stacks as is known in the art using standard techniques. The dimensions of minimum feature size can be made subnominal by the use of etch bias or OPC (Optical Proximity Correction). Therefore, if the minimum resolution of photo-definition is 130 nm, the etch bias can make the final line width 90 to 100 nm (standard practice in the art). Subnominal gate size is needed to accommodate the additional two gate electrodes for the transistors. These gate stacks will form two center P+ gate electrodes 40 of two transistors.

Referring to Fig. 6, a simple wet clean removes the residual gate dielectric 12. Then, a dielectric layer 22 is formed on the sides of the remaining P+ gate electrode 40 stacks by depositing a nitride layer up to, but preferably less than, about 2.0 nm in thickness. The dielectric layer 22 can alternatively be oxynitride or nitrided oxide. The dielectric layer 22 should be as thin as possible to still ensure proper insulation between the

gate electrodes 40 and 42a and 42b because of the possible formation of resistive regions. Resistive regions under the transistor gate will result in a channel region of lower conductivity and thus lower device performance. A light anisotropic nitride etch is used to clear the dielectric layer 22 from over the active areas, but keeps the dielectric layer 22 on the sidewalls of the central P+ gate electrode 40.

Referring to Fig. 7, the wafer is wet cleaned and prepared for the growth of a second gate dielectric 13. The second gate dielectric 13 is now grown over the substrate to a thickness substantially equal to or thinner than that of the original gate dielectric 12, which is still beneath the P+ polysilicon of the P+ gate electrode 40. This regrown gate dielectric 13 serves as the barrier between the outer gate electrodes 42a and 42b and the substrate 10.

Next, another polysilicon layer, having an N+ conductivity type, is deposited over the dielectric layer 22 and the newly formed gate dielectric 13. This will eventually form the two outer N+ gate electrodes 42a, 42b for the transistors. This N+ polysilicon layer can be up to about 50 nm thick, preferably 20 to 25 nm thick, and can be deposited in a similar manner as the first P+ polysilicon layer. The structure is subjected to anisotropic etching to remove a portion of N+ polysilicon layer and leave sidewalls on the dielectric layer 22. The N+ polysilicon layer is thus removed from over the substrate 10 except for the portion that remains on the sides of the dielectric layer 22 adjacent to the P+ gate electrode 40. The resulting structure shown in Fig. 7 consists of two free-standing structures having three gate electrodes 40, 42a and 42b.

Next, referring to Fig. 8, a conductive layer, preferably polysilicon, is deposited over each of the two three-gate structures (100 nm), followed by the masking and etching of this layer to leave a conductive cap 26 (e.g., a strap contact) in electrical contact with all

three gate electrodes 40, 42a and 42b of each structure. Alternatively, salicide formation of cap 26 can be used to save a masking step.

If forming the alternative embodiment illustrated in Fig. 2a, the conductive cap 26 described in reference to Fig. 8 can be patterned and etched to isolate separate conductive caps 26a and 26b over the three adjacent gate electrodes 40, 42a, and 42b. Alternatively, the outer gate electrodes 42a and 42b can be connected by a sidewall electrical contact (not shown).

Next, referring to Fig. 9, an insulating layer, preferably oxide or nitride, is formed over the structures (10-20 nm) and dry etched to form insulating sidewall spacers 28. Then source/drain regions 32 are formed by ion implantation 30. If the alternative embodiment illustrated in Fig. 2a is to be formed, the insulating layer should be formed between the separate conductive caps 26a and 26b to electrically isolate them from one another. At this point the device according to the invention is substantially complete, only to be followed by standard semiconductor processing, including the possible formation of capacitors and interconnect lines, or other devices as appropriate for the intended transistor function, be it as part of a DRAM memory cell, or otherwise.

For alternative embodiments, various other materials may be used for the gate electrodes other than polysilicon. Referring to Fig. 10a, silicon-germanium may be used to tailor the workfunctions of the gate electrodes. This may be appropriate if the difference in V_t between the P+ polysilicon and N+ polysilicon is excessive in relation to the power supply to the device. For low voltage applications, even a few hundred mV can be excessive. For memory access device applications, for instance, plus or minus 200 mV can cause the device to fail a margins test. A Si/Ge gate can allow adjustment of the workfunction up to 0.46 volts for 100% Ge. The Si/Ge material can be used to replace the P+ polysilicon. Referring to Fig. 10b, silicon carbide or silicon oxycarbide can also be used,

for either gate types with either P-type or N-type dopings. These compounds have larger band gap energies and as a consequence, lower electron affinities and different workfunctions in comparison to silicon. Various other crystalline structures with very small grain size and quantum confinement will also produce different workfunctions and may be utilized as is known in the art.

FIG. 11 illustrates a processor-based system (e.g., a computer system), with which semiconductor transistors constructed as described above may be used. The processor-based system comprises a central processing unit (CPU) 102, a memory circuit 104, and an input/output device (I/O) 100. The memory circuit 104 may be formed as one or more memory modules, each containing one or more integrated memory devices (e.g., DRAM devices) including transistor devices constructed in accordance with the invention. Also, the CPU 102 may itself be an integrated processor which utilizes transistor devices constructed in accordance with the present invention, and both the CPU 102 and the memory circuit 104 may be integrated on a single chip.

The above description and accompanying drawings are only illustrative of exemplary embodiments, which can achieve the features and advantages of the present invention. It is not intended that the invention be limited to the embodiments shown and described in detail herein. The invention can be modified to incorporate any number of variations, alterations, substitutions or equivalent arrangements not heretofore described, but which are commensurate with the spirit and scope of the invention. The invention is only limited by the scope of the following claims.

What is claimed as new and desired to be protected by Letters Patent of the United States is: